U.S. Department
of Transportation
**National Highway
Traffic Safety
Administration**

NHTSA

DOT HS 812 688

April 2019

Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs

DISCLAIMER

Suggested APA Format Citation:

Zhang, F., Subramanian, R., Chen, C.-L., & Noh, E. Y. (2019, April). *Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs* (Report No. DOT HS 812 688). Washington, DC: National Highway Traffic Safety Administration.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT HS 812 688		

4. Title and Subtitle	5. Report Date
Crash Report Sampling System: Design Overview, Analytic Guidance, and FAQs	April 2019
	6. Performing Organization Code: NSA-210
7. Author(s): Fan Zhang, Rajesh Subramanian, Chou-Lin Chen, Eun Young Noh	8. Performing Organization Report No.
9. Performing Organization Name: Mathematical Analysis Division, National Center for Statistics and Analysis, National Highway Traffic Safety Administration, 1200 New Jersey Avenue SE, Washington, DC 20590	10. Work Unit No. (TRAIS)
	11. Contract or Grant No.
12. Sponsoring Agency Name and Address: Mathematical Analysis Division, National Center for Statistics and Analysis, National Highway Traffic Safety Administration, 1200 New Jersey Avenue SE, Washington, DC 20590	13. Type of Report and Period Covered: NHTSA Technical Report
	14. Sponsoring Agency Code

15. Supplementary Notes

This document is an amended update of NHTSA's technical report under the same title (Report No. DOT HS 812 509) as a result of 2016 CRSS data revision.

The authors would like to thank Phillip Kott for his consultation.

Abstract

This report describes the Crash Report Sampling System (CRSS) sample design and weighting procedures, explains some basic concepts about estimation based on complex survey data, discusses issues of CRSS data analysis, and provides examples.

17. Key Words	18. Distribution Statement
NHTSA, CRSS, GES, NASS, sample design, complex survey data analysis, analytic guidance.	This document is available from the National Technical Information Service www.ntis.gov.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	30	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Acronyms

- CDS – Crashworthiness Data System
- CISS – Crash Investigation Sampling System – a replacement of CDS
- CRSS – Crash Report Sampling System – a replacement of GES
- FARS – Fatality Analysis Reporting System
- GES – General Estimates System
- MOS – measure of size
- NASS – National Automotive Sampling System
- NHTSA – National Highway Traffic Safety Administration
- PAR – police crash/accident report
- PJ – police jurisdiction
- PSU – primary sampling unit
- SSU – secondary sampling unit
- TSU – tertiary sampling unit

TABLE OF CONTENTS

1. Introduction

The National Highway Traffic Safety Administration developed and implemented the National Automotive Sampling System in the 1970s to make estimates of the motor vehicle crash experience in the United States. In 1988 NHTSA split the NASS into two surveys, the General Estimates System and the Crashworthiness Data System. Since then the same data collection sites have been used for GES data collection and a sub-sample of the GES sites has been used for CDS data collection. Given the shifts in population and the vehicle fleet, and the changing analytic needs of the safety community, the United States Congress authorized NHTSA to modernize its crash data collection system.

NHTSA implemented two new annual surveys, the Crash Report Sampling System, which replaced the GES, and the Crash Investigation Sampling System, which replaced the CDS.

This document provides an overview of the CRSS sample design (Chapter 2) and weighting procedure (Chapter 3). Sample design and weighting procedure determine the design options when CRSS data is analyzed using complex survey data analysis software.

Chapter 4 discusses some basic concepts on the analysis of complex survey data that justify the practice of using finite population point estimates and design variance estimates to make inference about model parameters. In Chapter 5 we provide examples to show how to make estimates using CRSS and GES data and discuss issues related to CRSS data analysis.

Finally, Chapter 6 catalogs and answers frequently asked questions on sampling and estimation of GES/CRSS.

While this report provides a broad overview of the design of CRSS, a supplemental NHTSA technical report, *Crash Record Sampling System: Sample Design and Weighting*, to be published by NHTSA this year illustrates the CRSS sample design and weighting procedures in greater detail.

2. CRSS Sample Design

CRSS was designed independent of other NHTSA surveys. The target population for the CRSS is the same as that for the GES, all police-reported motor vehicle crashes on trafficways. Because a nationwide direct selection of police crash reports is currently not feasible, the CRSS sample was selected in multiple stages to produce a nationally representative probability sample.

At the first stage the CRSS sample selection, 3,117 counties in the United States were grouped into 707 primary sampling units. A PSU in the CRSS is either a county or a group of counties. U.S. territories, some remote counties in Alaska, and small islands of Hawaii were excluded.

The 707 PSUs were stratified into 50 strata by the four Census regions, urban/rural, vehicle miles traveled, total number of crashes, total truck miles traveled, and road miles. Each of the 707 PSUs in the frame was assigned a measure of size equal to the combination of its estimated nine types of crash counts defined in Table 1 below. One large PSU was selected with certainty. From each of 50 PSU strata, 2 PSUs were selected by a stratified probability proportional to size (PPS) sampling method. This resulted in a sample of 101 PSUs. Then a sequence of sub-samples was selected from the 101 PSU sample with decreasing sample sizes, and in this process the PSU strata were collapsed if necessary. This process produced a sequence of nested PSU samples. These nested PSU samples allow NHTSA to change the PSU sample size without reselecting the sample in the future. Therefore, the final PSU sample is the result of a multiphase sampling, and the PSU sample selection probability is still approximately PPS.

For the 2016 CRSS, the 60 PSU sample was used: 59 PSUs from 26 PSU strata plus 1 certainty PSU. Among them, 7 PSUs refused to cooperate, so data was collected from 53 PSUs. A PSU level non-response adjustment was applied to mitigate the potential non-response bias. Due to the PSU non-responses, some PSU strata were collapsed for variance estimation resulting in 24 PSU strata and the certainty PSU was treated as a stratum. In the end, there were total 25 PSU strata for the 2016 CRSS. See Section 5 for more detail about PSU and PSU strata for variance estimation.

In the 2017 CRSS, 6 of the 7 non-responding PSUs were converted to responding PSUs and 1 replacement PSU was added. As a result, a total of 61 PSUs were selected and 60 PSUs cooperated. The number of PSU strata remained 25, the same as the 2016 CRSS.

The secondary sampling units were police jurisdictions. Each PJ in the selected PSUs was assigned a measure of size equal to a combination of crash counts in six categories of interest. PJs were then stratified into three PJ strata by their measure of sizes in each selected PSU. The Pareto sampling method (Rosén, 1997) was used to select PJ samples from each PJ stratum. The Pareto sampling method produces overlapping samples when a new sample is selected because of PJ frame change. This reduces the changes to the existing PJ sample. The PJ inclusion probability under the Pareto sampling is approximately PPS (Rosén, 1997). Across the 53 responding PSUs, a total of 350 PJs were selected and 337 PJs cooperated in 2016. For the 2017 CRSS, 397 PJs were selected and 393 PJs were cooperated. Weight adjustments were made to mitigate the potential bias caused by the non-responding PJs.

The tertiary sampling units were PARs. CRSS data collectors periodically receive new PARs from the selected PJs. All new PARs were sequentially stratified in the order they became available into nine PAR strata (see Table 1 below). These nine PAR strata were formed based on the results of NHTSA's internal data needs and public data needs studies. The PAR stratification was used to over-sample the following analysis domains so that enough cases could be selected from them:

- Crashes involving killed or injured pedestrians;
- Crashes involving killed or injured motorcycle occupants;
- Crashes involving killed or injured occupants in a late model year passenger vehicle; and
- Crashes involving killed or severely injured occupants in a non-late model year passenger vehicle.

From each PAR stratum, PAR sample was selected using the systematic sampling method. The sampling intervals were determined in such a way that predetermined target sample allocation by PAR strata was achieved and the final weights were approximately equal for all the PARs in the same PAR stratum. Equal weights reduce the sampling variance for the domain estimates. The target PAR sample size is around 50,000 every year. See the upcoming report, *Crash Record Sampling System: Sample Design and Weighting*, for more detailed information on CRSS sample design.

From each sampled PAR, approximately 120 data items about crash, event, vehicle, and people were coded. See *Crash Report Sampling System: Analytical User's Manual 2016*[1] for more information about data items coded in CRSS data files.

In the CRSS, the missing values (missing entries coded as "unknowns" and "not reported") in 27 selected variables were imputed by one of the following imputation methods:

- The sequential regression multivariate imputation (SRMI) method (Raghunathan et al, 2001).
- The univariate imputation method.
- The logical imputation.

These are the same imputation methods used for the GES. See Herbert (in press) for more detailed information on how these methods were used to impute missing items in the CRSS.

[1] National Highway Traffic Safety Administration. (2018, March). *Crash Report Sampling System analytical user's manual 2016* (Report No. DOT HS 812 510). Washington, DC: Author. Available at https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812510

Table 1: 2016 CRSS PAR Strata, Sample Allocation, and Estimated Crash Distribution

PAR Strata	Description	Target Sample Allocation (%)	Unweighted Distribution		Weighted Distribution	
			Resulting Sample Allocation (%)	Standard Error	Crash Distribution (%)	Standard Error
2	Crashes not in Stratum 1 that: • Involves a killed or injured (includes injury severity unknown) non-motorist	9	8.29	0.13	2.32	0.23
3	Crashes not in Stratum 1 or 2 that: • Involves a killed or injured (includes injury severity unknown) motorcycle or moped rider	6	4.97	0.11	1.35	0.08
4	Crashes not in Stratum 1-3 that: • At least one occupant of a late model year passenger vehicle is killed or incapacitated	4	2.74	0.08	0.41	0.03
5	Crashes not in Stratum 1-4 that: • At least one occupant of an older passenger vehicle is killed or incapacitated	7	6.12	0.11	1.43	0.10
6	Crashes not in Stratum 1-5 that: • at least one occupant of a late model year passenger vehicle is injured (including injury severity unknown)	14	15.73	0.17	7.50	0.42
7	Crashes not in Stratum 1-6 that: • involved at least one medium or heavy truck or bus (includes school bus, transit bus, and motor coach) with GVWR 10,000 lbs. or more	6	5.51	0.11	6.15	0.18
8	Crashes not in Stratum 1-7 that: • at least one occupant of an older passenger vehicle is injured (including injury severity unknown)	12	13.42	0.16	15.01	0.44
9	Crashes not in Stratum 1-8 that: • involved at least one late model year passenger vehicle, AND • No person in the crash is killed or injured	22	22.59	0.19	28.77	0.90
10	Crashes not in Stratum 1-9: • This includes mostly property damage only (PDO) crashes involving a non-motorist, motorcycle, moped, and passenger vehicles that are not late model year and any crashes not classified in strata 1-9.	20	20.62	0.19	37.06	0.83

Source: Estimates are from 2016 CRSS data.

- Unweighted estimates were computed by ignoring the sample design and weights.
- Late model year vehicle: vehicle that is no more than 4 year old.
- Older vehicle: vehicle that is more than 4 year old.

3. CRSS Weighting Procedures

The CRSS sample is the result of probability sampling featuring stratification, clustering, and selection with unequal probabilities. Because of these complex design features, the CRSS sample is not a simple random sample. Users need to use proper weights to produce unbiased and robust estimates. The CRSS weights are created with the following steps:

- Calculate the base weights (the inverse of selection probabilities) at all three stages (PSU, PJ, and PAR).
- Adjust the base weights for non-response[2] at all three stages to correct potential non-response bias.
- Adjust the weights for duplicate PARs.
- Calibrate the PJ and the PAR weights using the PSU level total PAR stratum counts to further correct potential non-response bias and coverage bias.
- Calibrate case weights by benchmarking Census resident population counts and FARS crash counts.

The final case weight is named as WEIGHT in the CRSS analysis files. See the upcoming report, *Crash Record Sampling System: Sample Design and Weighting*, for more detailed information on the CRSS weighting procedure.

[2] Non-responding PARs are incomplete PARs or non-readable PARs. Non-responding PJs and PSUs are PJs and PSUs refused to cooperate.

4. Basic Concepts of Complex Survey Data Analysis

4.1 Model Parameter Estimation

In standard statistical theory, we often assume that the data generated by nature or by a laboratory experiment follows a stochastic model. The model parameter that indexes the underlining model is of interest and needs to be estimated. For example, consider fatal indicators $\{y_1, y_2, \dots, y_N\}$:

$$y_k = \begin{cases} 1, & fatal\ crash \\ 0, & nonfatal\ crash \end{cases}, \quad k = 1, 2, \dots N$$

observed from the N crashes reported in 2016. One may view these observations as outcomes of independent and identical Bernoulli trials indexed by model parameter θ:

$$y_k \sim Bern(\theta), \qquad k = 1, 2, \dots N$$

And use the maximum likelihood estimator:

$$\hat{\theta}_N = \frac{1}{N}\sum_{k=1}^{N} y_k$$

to estimate the model parameter θ. If this model is correct, $\hat{\theta}_N$ is unbiased with respect to the model for θ:

$$E_{Bern}(\hat{\theta}_N) = \frac{1}{N}\sum_{k=1}^{N} E_{Bern}(y_k) = \theta$$

with variance:

$$Var_{Bern}(\hat{\theta}_N) = \frac{1}{N^2}\sum_{k=1}^{N} Var_{Bern}(y_k) = \frac{\theta(1-\theta)}{N} = O(N^{-1}).$$

Here E_{Bern} and Var_{Bern} are the expectation and variance with respect to model $Bern(\theta)$. Notice when N is very large, the model variance $Var_{Bern}(\hat{\theta}_N)$ becomes very small.

4.2 Finite Population Parameter Estimation

In the previous section, the model parameter θ is estimated by:

$$\hat{\theta}_N = \frac{1}{N}\sum_{k=1}^{N} y_k.$$

However, the quantity $\hat{\theta}_N = \sum_{k=1}^{N} y_k/N$ itself is also of interest because it gives a snapshot of the nation's fatal crash proportion in 2016. Similar statistics include N (2016 total number of crashes) and $\sum_{k=1}^{N} y_k$ (2016 total number of fatal crashes), etc. In other words, in addition to model parameters, we may also be interested in the functions of a set of realized (fixed) values. For example, the collection of all realized 2016 crashes $U = \{u_1, u_2, \dots, u_N\}$ can be viewed as a finite population. The functions of the attributes of the finite population, such as $\hat{\theta}_N$, N, and $\sum_{k=1}^{N} y_k$ are called finite population parameters.

Unfortunately, it is often cost-prohibitive to observe all the units in the finite population. Instead, a probability sample is selected and observed to estimate the finite population parameters.

A probability sample s is a subset of the finite population U selected under a probability sampling design. The key role of the probability sampling design is to define a probability space on U so we can use the sample s to estimate and make inferences about the finite population parameters. Chapters 2 and 3 briefly described how a probability sample of PARs was selected from a finite population of PARs for CRSS data collection and how the final CRSS weights were calculated.

It should be noted that for various reasons, it is inevitable to use design features such as stratification, clustering, and unequal selection probabilities to select the probability sample. For example, cluster sampling was used because it was too costly to obtain all PARs in the US to directly select a PAR sample. PARs in important analysis domains were assigned larger selection probabilities to ensure enough sample sizes for analysis. Stratification was used at all stages to reduce the sampling variance and assign different selection probabilities. These design features might induce a stochastic dependence among the resulting observations and alter the original distribution. As a result, the final sample is not a simple random sample, and the sampled observations may no longer follow the same model as the population from which they were drawn.

Under a probability sampling design, every unit u_k in the finite population $U = \{u_1, u_2, \dots, u_N\}$ has a positive probability π_k of being selected into the sample s. Assume sample $s = \{u_1, u_2, \dots, u_n\}$ has fixed sample size $n \leq N$ and define the selection indicator as:

$$I_k = \begin{cases} 1, & if\ u_k\ is\ selected\ into\ s \\ 0, & otherwise \end{cases} \qquad (k = 1,2, \dots, N)$$

The inverse of the inclusion probability $w_k = 1/\pi_k$ can be used to construct design-based point estimators of finite population parameters (i.e., they are unbiased or nearly unbiased under the probability-sampling design). For example, let the fatal indicator y_k be an attribute observed from crash u_k, then

$$\hat{\theta}_n = \frac{1}{N} \sum_{u_k \in s} w_k y_k$$

is design unbiased for the 2016 fatality proportion: $\hat{\theta}_N = \sum_{k=1}^{N} y_k/N$:

$$E_D(\hat{\theta}_n) = E_D\left(\frac{1}{N}\sum_{u_k \in s} w_k y_k\right) = E_D\left(\frac{1}{N}\sum_{k=1}^{N} w_k I_k y_k\right) = \frac{1}{N}\sum_{k=1}^{N} y_k = \hat{\theta}_N$$

Here the expectation E_D is with respect to the probability space defined by the sampling design. The sampling/design variance of $\hat{\theta}_n$, $Var_D(\hat{\theta}_n)$, is the variance of estimator $\hat{\theta}_n$ under repeated probability sampling. $Var_D(\hat{\theta}_n)$ depends on both the estimator $\hat{\theta}_n$ and the sample design. It should be noted that the point estimator $\hat{\theta}_n$ is design unbiased for the finite population parameter $\hat{\theta}_N$ regardless of whether the model assumed to generate the finite population is true or not.

4.3 Two-Step Sampling Procedure

Combining the concepts in the two previous sections, survey data can be viewed as the result of the following two step sampling procedure (Hartley & Sielken, 1975):

- Step 1: A finite population U of size N is generated by an infinite super-population model ξ.
- Step 2: A probability sample s of size $n \leq N$ is selected from the finite population U.

That is:

$$Model\ \xi \xrightarrow{Generation} U = \{u_1, u_2, \ldots, u_N\} \xrightarrow{Selection} s = \{u_1, u_2, \ldots, u_n\}$$

Under this two-step sampling view, the design unbiased point estimator is not only an unbiased estimator of the finite population parameter $\hat{\theta}_N$ under the probability based design, but also an unbiased estimator of the super-population model parameter θ if the (assumed) model is correct:

$$E_{\xi D}(\hat{\theta}_n) = E_\xi[E_D(\hat{\theta}_n)] = E_\xi[\hat{\theta}_N] = \theta$$

Here the expectation $E_{\xi D}$ is with respect to the two-step sampling process: the data generation by the model and the sample selection by the sample design. The total variance of a design unbiased point estimator $\hat{\theta}_n$ under this two-step sampling view can be decomposed as:

$$Var_{\xi D}(\hat{\theta}_n) = E_\xi[Var_D(\hat{\theta}_n)] + Var_\xi[E_D(\hat{\theta}_n)]$$

Since $E_D(\hat{\theta}_n) = \hat{\theta}_N$ and $Var_\xi(\hat{\theta}_N) = O(N^{-1})$, therefore $Var_\xi[E_D(\hat{\theta}_n)] = Var_\xi[\hat{\theta}_N] = O(N^{-1})$. So, when the finite population size N is large, the second term on the right is negligible. Therefore, if $\widehat{var}_D(\hat{\theta}_n)$ is a design unbiased estimator of $Var_D(\hat{\theta}_n)$, then it can also serve as an approximate estimator of the total variance when N is large:

$$\widehat{var}_{\xi D}(\hat{\theta}_n) \approx \widehat{var}_D(\hat{\theta}_n)$$

In addition, if the PSU sample is selected with-replacement or approximately so (when the sampling rate is low as in the CRSS), the with-replacement design variance estimator also captures the variance with respect to the model (Binder & Roberts, 2009).

In summary a design unbiased or nearly design unbiased point estimator can be used to estimate the finite population parameter regardless if the super-population model is correct or not. If the super-population model is correctly specified and the finite population parameter is unbiased with respect to the model for the model parameter, then the design unbiased estimator can also be used to estimate the model parameter. The design unbiased variance estimator for the design unbiased point estimator not only can be used to estimate the design variance of the design unbiased estimator, but also can be used to estimate its total variance when the finite population size is large.

From now on we only consider design unbiased or approximately design unbiased point estimators and their design variance estimators.

4.4 Design-Unbiased Point Estimator

Probability sampling defines a probability space so that the inclusion probability π_k for each sampled unit k can be derived and its inverse $w_k = 1/\pi_k$ can be used to weight the data to obtain (approximately) design unbiased estimators. The design-unbiased point estimator is robust because it is unbiased for the finite population parameter whether the super-population model that generated the finite population is true or not.

Unweighted estimators, on the other hand, may incur severe bias. In Table 1 for example, the unweighted crash distribution by PAR strata estimated from the 2016 CRSS sample, which is simply the 2016 CRSS sample allocation to the PAR strata, is quite different from the weighted distribution, which is an unbiased estimate of the actual crash distribution of all 2016 crashes by PAR strata.

4.5 Design Variance Estimation

The impact of the sample design must be recognized when one estimates $Var_D(\hat{\theta}_n)$. In Table 1, the unweighted standard errors ignored weights and the sample design. The weighted standard errors take the sample design (including the weights) into account. Table 1 shows ignoring the sample design may cause severe bias to the standard error estimates too.

Estimation methods and computer software have been developed to estimate $Var_D(\hat{\theta}_n)$. Specialized procedures for complex survey data analysis, such as SAS SURVEY procedures and SUDAAN procedures, should be used for the CRSS data analysis along with proper design statements. Because of the small CRSS PSU sampling fractions, the with-replacement design option can be used for CRSS data analysis.

Different variance estimation methods (for example, the jackknife variance estimation method and the Taylor series method) can be used to estimate the standard errors of CRSS estimates. We choose to use jackknife variance estimation method because our simulation study indicates it produces less biased variance estimates for small domain estimates. See Wolter (2007) for more information about design variance estimation under a complex sample design.

5. Estimation Examples

The following examples demonstrate how to use SAS or SUDAAN to calculate CRSS estimates.

- Example 1: Single-year CRSS estimates.
- Example 2: Combining multiple years of GES and CRSS data.
- Example 3: Composite estimates.
- Example 4: Domain estimates.

5.1 Example 1: Single-Year CRSS Estimates

The following SAS and SAS-callable SUDAAN programs show how design options are specified to make single-year CRSS estimates with major outputs from the SAS (Table 2) and SUDAAN (Table 3). In these examples the input data file CRSS2016.ACCIDENT is the 2016 CRSS crash-level data file. The 2016 CRSS estimates are computed for the imputed maximum injury severity of crashes (MAXSEV_IM).

We choose the Jackknife variance estimation method as the variance estimation method in SAS and SAS-callable SUDAAN programs. This also implicitly assumes the PSUs were selected with replacement or with a low sampling rate as in our case. We let the software to generate the Jackknife replicate weights.

The variable *PSUSTRAT* defines the PSU strata for variance estimation. In the 2016 CRSS, 7 PSUs did not cooperate. This left some PSU strata with only 1 responding PSU. In the variable *PSUSTRAT*, these PSU strata were collapsed with other strata to ensure at least 2 PSUs per stratum for variance estimation. Also, at the CRSS PSU sampling stage, 1 PSU was selected with certainty because of its large number of crashes. A certainty PSU is in fact a stratum therefore it is treated as a stratum in *PSUSTRAT*. The variable *PSU_VAR* identifies sampled PSUs for variance estimation. The PJs selected in the certainty PSU are treated as PSUs in *PSU_VAR*. Because of this, the number of PSUs used for the analysis is more than the original PSU sample size.

```
/*SAS Example*/

PROC SURVEYFREQ DATA=CRSS2016.ACCIDENT VARMETHOD=JK;
     STRATA PSUSTRAT;
     CLUSTER PSU_VAR;
     TABLES MAXSEV_IM;
     WEIGHT WEIGHT;
     FORMAT MAXSEV_IM MAXSEV.;
     RUN;
```

Table 2: Single-year CRSS estimates - SAS Output:

IMPUTED MAXIMUM INJURY IN CRASH					
MAXSEV_IM	**Frequency**	**Weighted Frequency**	**Std Dev of Wgt Freq**	**Percent**	**Std Err of Percent**
No Injury	21,605	4,666,609	291,822	68.4140	1.0909
Possible Injury	11,205	1,210,780	90,294	17.7504	0.8988
Minor Injury	7,648	687,613	38,447	10.0806	0.5057
Serious Injury	4,785	183,120	12,306	2.6846	0.1758
Fatal	913	34,748	2,136	0.5094	0.0353
Injured, Unknown	338	34,795	13,910	0.5101	0.2080
Died Before Crash	2	126.44829	96.51874	0.0019	0.0014
No Person Involved	15	3,337	999.12452	0.0489	0.0141
Total	46,511	6,821,129	378,064	100.000	

```
/*SAS-Callable SUDAAN Example*/

PROC CROSSTAB DATA=CRSS2016.ACCIDENT DESIGN=JACKKNIFE NOTSORTED;
     NEST       PSUSTRAT PSU_VAR;
     WEIGHT     WEIGHT;
     TABLES     MAXSEV_IM;
     CLASS      MAXSEV_IM;
     SETENV     ROWWIDTH=12 COLWIDTH=12 LABWIDTH=12;
     PRINT      NSUM="SAMSIZE" WSUM="POPSIZE" SEWGT="POP SE"
                / NSUMFMT=F6.0 WSUMFMT=F8.0 SEWGTFMT=F8.0;
     RFORMAT    MAXSEV_IM MAXSEV.;
     RUN;
```

Table 3: Single year CRSS estimates – SAS-Callable SUDAAN Output:

		IMPUTED MAXIMUM INJURY IN CRASH				
		Total	No Injury	Possible Injury	Minor Injury	Serious Injury
	SAMSIZE	46511	21605	11205	7648	4785
	POPSIZE	6821129	4666609	1210780	687613	183120
	POP SE	378064	291822	90294	38447	12306

		IMPUTED MAXIMUM INJURY IN CRASH			
		Fatal	Injured, Unknown	Died before Crash	No Person Involved
	SAMSIZE	913	338	2	15
	POPSIZE	34748	34795	126	3337
	POP SE	2136	13910	97	999

5.2 Example 2: Combining Multiple Years of GES and CRSS Data

Combining multiple years of data allows us to make year-to-year comparisons as well as better estimates for small domains. In the following example, two years of GES data (2014 and 2015) are combined with the 2016 CRSS data. First, annual crash counts by crash severity are estimated. Then pairwise comparisons are made among the annual estimates. Both analyses are implemented by SAS callable SUDAAN procedures. The same approach is also applicable to combining multiple years of GES data with multiple years of CRSS data.

CRSS sample selection is independent from GES sample selection. To capture this independence, a new stratification variable *STUDY* (*STUDY*=1 for GES and *STUDY*=2 for CRSS) is created in the data step. Annual samples within GES or annual samples within CRSS are not independent samples because the same PSU and PJ samples are used for data collection. A domain (sub-population) identification variable *YEAR* is created to make year-to-year comparisons. The variable *YEAR* has three categories: 2014, 2015, and 2016. In the data step, the PSU identification variable, *PSU_VAR*, is also defined for GES data. The sampled PJs in two GES certainty PSUs (13 and 14) were treated as PSUs for variance estimation.

The SUDAAN CROSSTAB procedure produces the output (Table 4) for annual parameter and variance estimates at all levels of crash severity (*CRASH_SEV*). Notice variable *STUDY* is used as an extra stratification variable so that the PSU identification variable, *PSU_VAR*, is the third variable listed in the NEST statement (*PSULEV*=3). The SUDAAN DESCRIPT procedure produces the output (Table 5) for the pairwise comparisons between the annual estimates.

```
PROC FORMAT;
     VALUE SEVERITY 1="FATAL" 2="INJURY" 3="PDO";
     RUN;

DATA COMBINED;
     SET CRSS2016.ACCIDENT (IN=CRSS2016)
          GES2015.ACCIDENT (IN=GES2015)
          GES2014.ACCIDENT (IN=GES2014);
     STUDY = GES2014 + GES2015 + CRSS2016*2;
     YEAR = CRSS2016*2016 + GES2015*2015 + GES2014*2014;
     IF (GES2015 OR GES2014) THEN DO;
          IF PSUSTRAT IN (13, 14) THEN PSU_VAR=PJ;
          ELSE PSU_VAR=PSU;
     END;
     IF MAXSEV_IM=4 THEN CRASH_SEV=1; /*FATAL CRASH*/
     ELSE IF MAXSEV_IM IN (1,2,3,5) THEN CRASH_SEV=2; /*INJURY
     CRASHES*/
     ELSE IF MAXSEV_IM IN (0,6,8) THEN CRASH_SEV=3; /*PDO
     CRASHES*/
     RUN;

PROC CROSSTAB DATA=COMBINED FILETYPE=SAS DESIGN=JACKKNIFE
NOTSORTED;
     NEST       STUDY PSUSTRAT PSU_VAR / PSULEV=3;
     WEIGHT     WEIGHT;
     CLASS      YEAR CRASH_SEV;
     TABLES     YEAR*CRASH_SEV;
     SETENV     ROWWIDTH=20 COLWIDTH=20 LABWIDTH=40;
     PRINT      NSUM="SAMSIZE" WSUM="TOTAL" SEWGT="SE TOTAL"
                / NSUMFMT=F8.0 WSUMFMT=F10.0 SEWGTFMT=F9.0;
     RFORMAT    CRASH_SEV SEVERITY.;
     RTITLE     "GES 2014, 2015 and CRSS 2016 Crash Severity
Comparison";
     RUN;

PROC DESCRIPT DATA=COMBINED FILETYPE=SAS DESIGN=JACKKNIFE
NOTSORTED TOTALS;
     NEST       STUDY PSUSTRAT PSU_VAR / PSULEV=3;
     WEIGHT     WEIGHT;
     CLASS      YEAR CRASH_SEV;
     TABLES     CRASH_SEV;
     VAR        _ONE_;
     PAIRWISE   YEAR / NAME="YEAR TO YEAR COMPARISON";
     SETENV     ROWWIDTH=20 COLWIDTH=20 LABWIDTH=40;
     PRINT      NSUM="SAMSIZE" TOTAL="DIFF" SETOTAL="DIFF STE"
                LOWTOTAL UPTOTAL
                / NSUMFMT=F10.0 TOTALFMT=F12.0 SETOTALFMT=F12.0
```

```
                LOWTOTALFMT=F12.0 UPTOTALFMT=F12.0;
    RFORMAT    CRASH_SEV SEVERITY.;
    RUN;
```

Table 4: Crash Severity Estimates (PROC CROSSTAB)

Variance Estimation Method: Delete-1 Jackknife
GES 2014, 2015 and CRSS 2016 Crash Severity Comparison
by: Crash Date (Year), CRASH_SEV.

Crash Date (Year)		CRASH_SEV			
		Total	FATAL	INJURY	PDO
Total	SAMSIZE	156619	2762	80364	73493
	TOTAL	19164503	80298	5479428	13604777
	SE TOTAL	882869	3970	220727	713988
2014	SAMSIZE	53030	895	27447	24688
	TOTAL	6058524	24296	1647726	4386502
	SE TOTAL	404124	2572	94199	329222
2015	SAMSIZE	57078	954	28941	27183
	TOTAL	6284851	21255	1715394	4548203
	SE TOTAL	402086	1474	95838	328529
2016	SAMSIZE	46511	913	23976	21622
	TOTAL	6821129	34748	2116308	4670073
	SE TOTAL	378064	2136	116882	292144

Table 5: Pairwise Comparisons by Crash Severity (PROC DESCRIPT)

GES 2014, 2015 and CRSS 2016 Crash Severity Comparison
by: Variable, CRASH_SEV, Contrast.

		Contrast		
CRASH_SEV		YEAR TO YEAR COMPARISON: (2014,2015)	YEAR TO YEAR COMPARISON: (2014,2016)	YEAR TO YEAR COMPARISON: (2015,2016)
Total	SAMSIZE	110108	99541	103589
	DIFF	-226327	-762605	-536278
	DIFF STE	115987	553397	551911
	Lower 95% Limit Cntrst Total	-456901	-1862723	-1633440
	Upper 95% Limit Cntrst Total	4246	337512	560884
FATAL	SAMSIZE	1849	1808	1867
	DIFF	3041	-10452	-13493
	DIFF STE	2526	3343	2595
	Lower 95% Limit Cntrst Total	-1979	-17098	-18652
	Upper 95% Limit Cntrst Total	8062	-3806	-8335
INJURY	SAMSIZE	56388	51423	52917
	DIFF	-67668	-468582	-400914
	DIFF STE	32521	150117	151150
	Lower 95% Limit Cntrst Total	-132317	-767004	-701391
	Upper 95% Limit Cntrst Total	-3018	-170160	-100437
PDO	SAMSIZE	51871	46310	48805
	DIFF	-161701	-283572	-121871
	DIFF STE	90587	440154	439635
	Lower 95% Limit Cntrst Total	-341782	-1158568	-995837
	Upper 95% Limit Cntrst Total	18380	591425	752096

It should be noted the difference between a CRSS and GES estimate may be compounded by the actual difference, sampling errors, and the potential bias of GES estimate. For example, NHTSA had noticed GES fatal crash estimate is underestimated. The 2014, 2015, and 2016 FARS fatal crash counts are 30,056, 32,166, and 34,439 respectively (see NHTSA's *Traffic Safety Facts 2015* and *Traffic Safety Facts 2016*). Table 4 shows that the 2014 and 2015 GES fatal crash estimates (24,296 with standard error 2,572 and 21,255 with standard error 1,474) are significantly lower than their FARS counterparts. In table 5, the significant difference between the 2016 CRSS fatal estimate and the 2015 GES fatal estimate (34,748 – 21,255 = 13,493 with standard error for the difference (DIFF STE) 2,595) contains the real difference, sampling errors, and the bias of GES. Different variables may suffer different degrees of bias. However, the comparison among GES estimates are less likely affected by the bias.

In Table 5 the standard errors for the differences (DIFF STE) between GES 2014 and GES 2015 are much smaller than those between GES and CRSS annual estimates (GES 2014 vs. CRSS 2016 or GES 2015 versus CRSS 2016) because GES annual estimates are positively correlated while the GES estimate is independent from the CRSS estimate.

5.3 Example 3: Composite Estimates

A composite estimate refers to a function of estimates made from different samples. FARS is an annual census survey of all fatal crashes. FARS data collection is independent from GES or CRSS data collection. FARS's target population – fatal crashes – is a sub-population of GES or CRSS target population. FARS finite population counts do not have design variance because all fatal crashes are observed. Therefore, FARS data can provide very good estimates for the fatal domain (sub-population). Because of this, it is sensible to make a composite estimate using both FARS count and CRSS estimate.

In the following example, a composite estimate of the total number of persons not in motor vehicles but involved in a crash is made in the following steps:

- Calculate the count of persons not in motor vehicles from fatal crashes in FARS: $\hat{\theta}_{FARS}$.
- Estimate of the number of persons not in motor vehicles from non-fatal crashes in CRSS: $\hat{\theta}_{CRSS}$.
- Create a composite estimate for total number of persons not in motor vehicles: $\hat{\theta}_C = \hat{\theta}_{FARS} + \hat{\theta}_{CRSS}$.

Specifically, to estimate the total number of persons involved in a crash but not in motor vehicles in 2016, we first aggregate the variable *PEDS* (Number of Persons Not in Motor Vehicles) in the 2016 FARS accident file over all 2016 fatal crashes to calculate the number of people not in motor vehicles in the FARS fatal crashes: $\hat{\theta}_{FARS}$ =7,661. Then we use the following SAS codes to estimate the number of people not in motor vehicles in the non-fatal crashes from the 2016 CRSS: $\hat{\theta}_{CRSS}$=164,682. Table 6 provides the CRSS estimates.

```
PROC FORMAT;
     VALUE FATAL 1="FATAL" 0="NON-FATAL";
     RUN;
```

```
DATA CRSS2016;
     SET CRSS2016.ACCIDENT;
     FATAL=(MAXSEV_IM=4); /*FATAL=1 for FATAL CRASHES*/
     RUN;

PROC SURVEYMEANS DATA=CRSS2016 VARMETHOD=JK SUM;
     STRATA PSUSTRAT;
     CLUSTER PSU_VAR;
     DOMAIN FATAL;
     VAR    PEDS;
     WEIGHT WEIGHT;
     FORMAT FATAL FATAL.;
     RUN;
```

Table 6: Estimate of Persons Not in Motor Vehicle in CRSS

Statistics		
Variable	**Sum**	**Std Dev**
PEDS	172,598	16,484

Domain Analysis: FATAL			
FATAL	**Variable**	**Sum**	**Std Dev**
NON-FATAL	**PEDS**	164,682	15,864
FATAL	**PEDS**	7,915.809341	948.120685

Finally, the composite estimate of the total number of persons not in motor vehicles in 2016 is estimated by combining the FARS count and the CRSS estimate:

$$\hat{\theta}_C = 7{,}661 + 164{,}682 = 172{,}343$$

The estimated standard error of $\hat{\theta}_C$ is:

$$se(\hat{\theta}_C) = \sqrt{Var(\hat{\theta}_{FARS}) + Var(\hat{\theta}_{CRSS})}$$
$$= se(\hat{\theta}_{CRSS}) = 15{,}864$$

Covariance of $\hat{\theta}_{FARS}$ and $\hat{\theta}_{CRSS}$ is zero because $\hat{\theta}_{FARS}$ and $\hat{\theta}_{CRSS}$ are independent. $Var(\hat{\theta}_{FARS}) = 0$ because FARS is a census, and $\text{se}(\hat{\theta}_{CRSS}) = 15{,}864$ from Table 6.

It should be noted that in this example the domain estimate $\hat{\theta}_{CRSS}$ (total persons not in motor vehicles among non-fatal crashes in CRSS) should be estimated from the full CRSS sample although we only need the estimate of persons not in motor vehicles among the non-fatal crashes. Sub-setting the full sample for domain estimation may produce a biased variance estimate (see Graubard & Korn, 1996).

5.4 Example 4: Domain Estimates

Domain estimate refers to the statistics for a subpopulation. It is important to use the full sample for domain estimation. It may produce biased variance estimate by sub-setting the full sample for domain estimation.

In SAS SURVEY procedures, domains are specified by the variables listed in the TABLES and/or the DOMAIN statement. The SAS BY statement sub-sets the full sample for one domain at a time, therefore it should not be used to produce domain estimates. In SAS-callable SUDAAN procedures, domains are specified by the variables listed in the TABLES and/or the SUBPOPN statement.

The following SAS program estimates the percentage of alcohol-involved non-fatal crashes. To this end, we defined a variable *FATAL* in the SAS data step and classify all crashes into two domains: fatal and non-fatal crashes. The domains were defined by variable *FATAL* in the DOMAIN statement.

```
PROC FORMAT;
     VALUE FATAL 1="FATAL" 0="NON-FATAL";
     RUN;

DATA CRSS2016;
     SET CRSS2016.ACCIDENT;
     FATAL=(MAXSEV_IM=4); /*FATAL=1 FOR FATAL CRASHES*/
     ALCOHOL=(ALCHL_IM=1); /*ALCOHOL=1 IF ALCOHOL INVOLVED*/
     RUN;

PROC SURVEYMEANS DATA=CRSS2016 VARMETHOD=JK MEAN SUM CLM;
     STRATA    PSUSTRAT;
     CLUSTER   PSU_VAR;
     WEIGHT    WEIGHT;
     VAR       ALCOHOL;
     DOMAIN    FATAL;
     FORMAT    FATAL FATAL.;
     RUN;
```

Table 7: SAS PROC SURVEYMEANS domain estimates

Domain Analysis: FATAL							
FATAL	**Variable**	**Mean**	**Std Error of Mean**	**95% CL for Mean**		**Sum**	**Std Dev**
NON-FATAL	**ALCOHOL**	0.047271	0.002984	0.04117653	0.05336488	320797	22254
FATAL	**ALCOHOL**	0.270563	0.019973	0.22977126	0.31135389	9401.508462	765.604811

The following SAS-callable SUDAAN program also estimates the percentage of alcohol involved crashes by two domains: fatal and non-fatal crashes. The domains were defined by variable *FATAL* in the TABLES statement.

```
PROC DESCRIPT DATA=CRSS2016 DESIGN=JACKKNIFE NOTSORTED;
     NEST       PSUSTRAT PSU_VAR;
     WEIGHT     WEIGHT;
     CLASS      FATAL;
     TABLES     FATAL;
     VAR        ALCOHOL;
     SETENV     ROWWIDTH=15 COLWIDTH=15 LABWIDTH=15;
     PRINT      NSUM="SAMSIZE" WSUM="POPSIZE" MEAN="MEAN"
                SEMEAN="MEAN SE" LOWMEAN UPMEAN
                / NSUMFMT=F8.0 WSUMFMT=F10.0 MEANFMT=F6.4
                SEMEANFMT=F6.4 LOWMEANFMT=F6.4 UPMEANFMT=F6.4;
     RUN;
```

Table 8: SUDAAN domain estimates

Variable		FATAL		
		Total	0	1
ALCOHOL	SAMSIZE	46511	45598	913
	POPSIZE	6821129	6786381	34748
	MEAN	0.0484	0.0473	0.2706
	MEAN SE	0.0030	0.0030	0.0200
	Lower 95% Limit Mean	0.0423	0.0412	0.2298
	Upper 95% Limit Mean	0.0546	0.0534	0.3114

6. Frequently Asked Questions

1. **What is CRSS?**
   A. The Crash Report Sampling System is NHTSA's new national probability-based crash sampling system designed to replace the General Estimates System.
2. **What data does CRSS collect and what does it represent?**
   A. CRSS samples police crash reports and codes the information into a data file. The CRSS data, when used with the accompanying weights, are nationally representative of all police-reported motor vehicle traffic crashes where the first harmful event occurred on a public trafficway.
3. **When did NHTSA transition from GES to CRSS?**
   A. The year 2015 was the last year of data collection through GES. CRSS was designed and implemented over a multi-year effort and started collecting data in January 2016.
4. **Why did NHTSA transition from GES to CRSS?**
   A. The Congress directed and provided funds to NHTSA to modernize its data collection system. The GES had used the same data collection sites since 1988. Over time, the population has shifted, the vehicle fleet and the analytic needs of the safety community have changed. In addition, the existing GES police jurisdiction samples and weights became outdated as the PJ population changed.
5. **How is CRSS different from GES in terms of its sample design?**
   A. The following are some major differences in sample designs of CRSS and GES:
      - Independent sample: The CRSS sample design is independent from the GES or any other NHTSA's surveys, including NHTSA's new Crash Investigation Sampling System that replaces the NASS Crashworthiness Data System. In comparison, the GES and the CDS samples were nested, i.e., the CDS used a subset of the GES data collection sites. The independent design allows NHTSA to optimize each system, CRSS and CISS.
      - Different formation of PSUs: In both CRSS and GES, a PSU is either a county or a group of counties. In CRSS, the nation was partitioned into 707 PSUs, while in GES 1195 PSUs were formed. CRSS's average PSU size is bigger than GES. This resulted in more equal weights in CRSS. In addition, a new composite PSU measure of size variable using various estimated crash counts was used in CRSS.
      - Finer PSU stratification: In the GES, 60 PSUs were selected from 12 PSU strata formed by census region and urbanicity type. In the CRSS, 60 PSUs were selected from 25 PSU strata formed by census region, urbanicity type, vehicle miles traveled, total number of crashes, total truck miles traveled, and road miles.
      - Scalable PSU sample: The CRSS PSU sample size can be increased without changes to the existing PSU sample while the corresponding selection probabilities are still trackable. This enables NHTSA to accommodate potential budget fluctuations with minimum operational costs and efforts.
      - Scalable PJ sample: The second stage sampling frame, the police jurisdictions in the selected PSUs, changes over time. Consequently, the PJ sample needs to be reselected occasionally to maintain adequate sample size or to cover the updated PJ frame. The Pareto sampling method is used for the PJ sample selection to reduce the changes to the existing PJ sample when a new PJ sample is selected.

- Alignment with data needs: The PAR strata were revised based on data needs. The PAR stratification is used to oversample the following analysis domains so that enough cases can be selected from those strata:
  - Crashes involving killed or injured pedestrian;
  - Crashes involving killed or injured motorcycle occupant;
  - Crashes involving killed or injured occupant of late model year passenger vehicle;
  - Crashes involving killed or severely injured occupant of non-late model year passenger vehicle.

- Optimized sample allocation: CRSS PSU, PJ, and PAR sample sizes were determined by minimizing the variance of a simplified variance estimator subject to fixed cost.
- Flexible system to allow mid-year changes: CRSS allows mid-year changes of sampling parameters such as PAR sampling intervals, PJ sample changes etc. to cope with unanticipated changes. GES sample parameters could not be changed in the middle of the year.
- Weight Adjustment: In the CRSS, weights of non-responding PSUs, PJs, and PARs, and duplicate PARs are adjusted to mitigate potential bias. In addition, case weights are calibrated by benchmarking Census resident population counts and FARS crash counts.

6. How is CRSS similar to GES in sample design?

A. The following are some major common features between the sample designs of CRSS and GES:
- The CRSS target population is the same as the GES, i.e., all police-reported crashes of motor vehicles occurring on a public trafficway.
- Both CRSS and GES collect information from police crash reports.
- Both surveys have a three-stage sample design: PSU, PJ, and PAR sample selection.
- In both surveys, PSUs and PJs have selection probabilities proportional to their measure of sizes.
- In both surveys, PAR samples are selected using systematic sampling.
- Both surveys tried to achieve equal-weights within PAR stratum.

7. How do the CRSS analysis files (data sets) differ from the GES?

A. The CRSS analysis files are almost the same as the GES analysis files. They have the same variables with the same names except CRSS no longer codes the land use variable. A new PSU identification variable *PSU_VAR* was created in CRSS for variance estimation (see Example 1).

8. Is the difference in total crashes (or other metric of choice) between the 2015 GES and the 2016 CRSS estimates due to the design change or an actual increase?

A. The difference between the 2015 GES total crash estimate and 2016 CRSS total crash estimate may result from the following:
- The actual difference between the 2015 and 2016 total crash counts;
- The sampling errors of 2015 GES and 2016 CRSS total crash estimates;
- The potential bias of 2015 GES total crash estimate.

For example, in the past we have observed GES fatal crash underestimation. When a CRSS estimate is compared to a biased GES estimate, the observed difference is confounded by the GES bias. However, different variables suffered different degrees of bias. The

comparisons among GES estimates are less likely affected by the bias. See Example 2 for more discussion on this.

9. **Is the difference between the 2015 GES total crash estimate and the 2016 CRSS total crash estimate statistically significant?**
   A. No. The difference between the 2016 CRSS total crash estimate and the 2015 GES estimate is not significantly different from zero due to the large variance of the difference estimate. It should be noted this only means given the data we have, we do not have evidence to reject the null hypothesis that the two years have the same total number of crashes. This does not necessarily mean the null hypothesis is true. See Example 2 for more detailed information about this test and computer codes.
10. **Can we combine FARS and CRSS data for estimates and analysis?**
   A. Yes. See Example 3, Chapter 5.
11. **How should data users compute the variance for CRSS estimates?**
   A. CRSS sample is the result of complex survey sampling, and therefore is not a simple random sample. Software specialized in complex survey data analysis such as SAS SURVEY procedures or SUDAAN procedures should be used to make estimates from CRSS sample. Using these specialized softwares along with the appropriate design and weight statements, the sampling variance can be estimated. Failing to take the sample design and weights into account in estimation may incur severe bias to the point and variance estimates. See Chapters 4 and 5 for some basic concept of complex survey data analysis, and SAS and SUDAAN example programs on how to estimate the variances for CRSS estimates.
12. **Are there issues of bias with GES estimates?**
   A. Yes. As we have seen in the past, GES fatal crash estimates and fatality estimates were significantly lower than the corresponding FARS counts. Therefore, NHTSA has been using FARS data for fatal crash counts. Different variables may suffer different degrees of bias. But trends within GES are less likely to be affected by the bias.
13. **If GES underestimated fatal crashes, how has this issue been addressed in CRSS? Will it still be necessary to replace CRSS fatal crashes with FARS data, or would CRSS be able to stand alone for all analyses of fatal/non-fatal crashes?**
   A. In the CRSS weighting procedure, the fatal estimate was calibrated to FARS fatal count. Therefore, the CRSS fatal estimate matches with the FARS fatal crash counts. CRSS can be used to make fatal estimates, but using FARS data is recommended as FARS is a nationwide census of all fatal crashes. NHTSA will continue to use FARS for fatality counts and CRSS for injury and PDO related estimates.
14. **Is there anything different need to do with CRSS data in producing estimates for very small sample size?**
   A. The problem associated with a small sample size does not change from GES to CRSS. See Rao & Molina 2015 for more details on small area estimation.
15. **How are missing data addressed in CRSS?**
   A. As in GES, key CRSS variables with missing information are imputed first using the sequential regression multivariate imputation procedure and then using the simple random sampling with-replacement imputation method and the logical imputation method. See Herbert (in press) for more details.

7. References

Binder, D. A., & Roberts, G. (2009). *Design- and model-based inference for model parameters*. In D. Pfeffermann& C. R. Rao, eds. *Handbook of Statistics 29*, Vol. 29B, , Amsterdam; Heidelberg: Elsevier, North-Holland.

Graubard, B. I., & Korn, E. L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology, Vol. 144*, No. 1, pp 102-106.

Hartley, H. O., & Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics, Vol. 31*, No. 2, pp 411-422.

Herbert, G. C. (in press). *Crash Report Sampling System: Imputation*. Washington, DC: National Highway Traffic Safety Administration.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. New York: Wiley Series in Survey Methodology.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference, Vol. 62*, pp. 159-191.

Wolter, K. (2007). Introduction to variance estimation. New York: Springer-Verlag New York, Inc.

Zhang, F., Noh, E. Y., Subramanian, R., & Chen, C-L. (in press). *Crash Report Sampling System: Sample Design and Weighting*. Washington, DC: National Highway Traffic Safety Administration.

DOT HS 812 688
April 2019

U.S. Department
of Transportation

**National Highway
Traffic Safety
Administration**

14187-040219-v3a